Warsaw School of Economics
Collegium of Economic Analysis

Damian Przekop

Algorithmic construction of predictors for rare events modeling – application fraud example

Abstract of the PhD thesis prepared under the guidance of
Professor Marek Gruszczyński (supervisor)
and
Dr. Marcin Owczarczuk (co-supervisor)
Collegium of Economic Analysis
Institute of Econometrics

Warsaw 2017

1. Aim of the thesis

In this thesis, the novel methodology of rare events modeling, based on the anti-fraud example, is presented. As the choice of statistical model used in the research improves the predictive capabilities of a solution only to some extent, most of the attention is paid to the choice of proper predictors.

The motivation in conducting this research was to verify whether the predictive power of a rare events model can be improved through adding new variables, formulated on the basis of proposed algorithms.

The main finding of the research is that model enrichment with additional predictors leads to the further improvement of the predictive power of a rare events model. The paper presents two algorithms as a solution to the problem of identifying the fraudulent intention of a customer. Their purpose is to generate variables that contribute to the improvement of fraud models' predictive power. The method presented is of general use as it produces variable input for every tool equipped with a variable-selection algorithm. The approach is based on the detection of uncommon behavior. Its performance is illustrated on a dataset from one of the Polish banks.

The thesis consists of two parts. The first is an exhaustive literature review. The second part is devoted to description of method and empirical study.

2. Topic of and motivation for the research

Despite the fact that fraud detection is a widely exploited topic in the research, the literature concerning loan fraud (i.e. application fraud) is very limited. Dorfleitner and Jahnes (2014) explain this state as a consequence of difficulty in obtaining data and censorship of the results obtained. Both factors aim to limit fraudsters' advantage in outsmarting bank security precautions. According to Hand (2007), banks almost always remain one step behind in this arms race.

There are three major components of risk that a bank faces: credit risk, operational risk, and market risk. Operational risk is the least identified and may account for up to 35% of the

volatility of universal banks' financial results (Cruz 2002). According to the Basel Committee, operational risk is *the risk of a change in value caused by the fact that actual losses, incurred for inadequate or failed internal processes, people and systems, or from external events (including legal risk), differ from the expected losses* (Basel Committee 2006). Basel III introduces the issue of liquidity risk, which is going to be strictly regulated from 2019 (Iwanicz-Drozdowska, 2012).

Application fraud is a part of operational risk defined by Basel II (Basel Committee 2006). It is committed by bank customers when applying for a loan. Sanusi et al. (2015) state that despite the fact that application fraud can account for more than half of the total amount of losses to which the banking sector is exposed, this phenomenon is still not identified well. One of the obstacles to achieving this goal are difficulties in making a clear distinction between application fraud cases (which are a part of operational risk) and cases of defaulted loans (which are a part of credit risk). It should be noted that in both cases the result is the same: an unrepaid loan. However, the motivations of the two types of customers remain different. This topic was investigated by Hartmann-Wendels et al. (2009), Mählmann (2010), and Dorfleitner and Jahnes (2014).

Banks build and implement anti-fraud systems as far as possible. These can consist of simple expert rules but also of advanced statistical models or even of social networks that make it possible to identify connections between the customers analyzed and previously committed frauds or known fraudsters. Fraudsters, over the course of time, develop more and more sophisticated ways of deceiving banks' security precautions, leading banks to tap into more and more advanced analytical tools that enable them to identify non-obvious fraudulent behavior.

According to annual surveys conducted among Polish financial institutions, the rate of fraud is increasing each year.[1] In each survey, a majority of respondents indicate that they expect it to increase. In 2012, 58% of respondents expected an increase in the rate of fraud, in 2013 60% and in 2014 nearly 75%.

---

[1] Report on fraud research on the financial market, which is a summary of research conducted by the Conference of Financial Companies and the EY audit company. 2015 Edition: https://kpf.pl/pliki/raporty/raport_naduzycia_2015.pdf (02.02.2017), 2016 Edition: https://kpf.pl/pliki/raporty/raport_naduzycia_2016.pdf (02.02.2017).

In 2015, 85% of respondents conducted investigations due to potential fraud events, 5% of them doing so more than 10 000 times. As a result, 43% of respondents reported a crime, 8% of them more than 1000 times.

In 2016, a majority of the respondents (60%) that were affected by fraud declared a total loss of up to 500 000 PLN. Every third institution declared a loss exceeding one million PLN. The two types of institutions mostly affected by instances of fraud are universal banks and leasing companies.

Application fraud constitutes 7% of all fraud in the Polish financial market.

The motivation to conduct this research is the deficit of modeling methodology that can be exploited in solving the problem of prediction of rare events that are dynamic and time-variant, such as application fraud.

3. Research problem

The subject of this research is the segment of *new* customers, i.e. those who do not have a product history with the bank in question. Despite the fact that the concentration of fraud among *new* customers is several times as high as among customers with a product history, cases of fraud make up only 3% of all loans granted in this segment. Keeping in mind the definition of an outlier proposed by Hawkins (1980) (*an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism*) and the nature of the phenomenon discussed, rare events of fraud occurrence can be perceived as outlying observations when compared to usual behavior of bank customers.

The issue under consideration is similar to rare events modeling, which is widely exploited in econometric literature. However, the discussion is narrowly focused on reducing the bias of estimators, rather than on explaining the variability of the phenomena discussed.

The solution proposed in the literature for improving fraud models' predictive power is the idea of a hybrid approach using information gained from the application of unsupervised learning algorithms in more global research, based on supervised learning methods. Such an

approach was applied in the works of Thornton et al. (2014), Favaresh and Sepehri (2011), and Krivko (2010). In these papers the researchers enriched their models with additional predictors, which led to the further improvement of the models' quality.

According to the research results obtained, unsupervised learning methods cannot compete with supervised learning algorithms, and additional information derived from their results does not contribute to the further improvement of the predictive power of models based on a supervised learning approach.

Finding the abnormality of selected observations requires in this case searching for specific transformation of features that are useful in solving a predictive problem. The first type of fraudster-specific predictor is a unique combination of features that is specific to fraudsters. The second type of fraudster-specific predictor is the relative value of a continuous variable within a defined peer group, which should be perceived here as a group of the most comparable customers.

4. Hypotheses

The thesis is that there exists a possibility to improve the quality of predictions delivered by predictive models as a result of model enrichment with additional predictors. These allow better understanding of rare event variation – in this case, application fraud.

- Hypothesis 1: rare events can be predicted more accurately using a new algorithm which creates predictors based on the idea of of distribution of continuous variables within a defined peer group.
- Hypothesis 2: rare events can be predicted more accurately using a new algorithm which creates predictors based on the concept of decision tree rules.
- Hypothesis 3: models enriched with the above-mentioned additional predictors are usually overtrained; this overtraining can be diminished by reduction of the number of predictors used, without any detriment to the predictive power of the new models.

5. The method of algorithmic construction of predictors

In the thesis two algorithms for predictors construction are presented. The new features which they produce improve the predictive power of rare event models.

The first algorithm generates variables that are unique combinations of features specific to fraudsters. The idea is based on the concept of rules generated by decision trees. The algorithm makes it possible to keep in one variable information that comes from more than one initial dataset variable. It allows the reduction of the degrees of freedom used in modeling.

Adding derived variables to the dataset used as the input for the model can be perceived as an ensemble classifier. Such an approach is termed in the literature a *multi-inducer* approach (Rokach, 2010). Joining two paradigms of classification allows one to obtain a synergy effect that can result in a more precise outcome than in the case of the solution's components.

The essence of the first algorithm is identification of a unique combination of features that can be connected to a rare event such as application fraud. Its construction allows the creation of variables that can consist of any possible combination of variables from the initial dataset. The algorithm makes it possible to capture more and less obvious predictors that represent fraud trends hidden in the data. In some fraud cases, only a combination of features may be useful in predicting an event.

The second algorithm is based on the logic of distribution of a continuous variable within a peer group defined by the values of selected nominal variables. This approach was developed by the author based on his professional experience of fraud detection.

The underpinning of the second algorithm refers to the concept of Peer Group Analysis proposed by Kim and Sohn (2012). A valuable piece of information in terms of solving predictive problems is not the behavior of the examined object itself but the expression of its unusualness. Such construction of variables allows the researcher to keep in one variable information that refers not only to a particular object but to the rest of population taken into account in the research.

The second algorithm aims to capture information helpful in fraud detection that can be seen only when interpreted using the concept of peer groups. The same absolute value of a

variable (e.g. salary) can mean two different things when interpreted compared to two different peer groups. The aim of the algorithm is to emphasize the meaning of the relative value of a variable.

6. Results

In order to verify the influence of the introduced algorithms on the improvement of anti-fraud models' predictive power, a regularized LASSO regression was used (as proposed by Tibshirani, 1996). The main advantage of this method is its applicability to multidimensional problems that include many predictors, even if the number of predictors exceeds the number of degrees of freedom available. This property is one of the most crucial in solving this problem.

The experiment applies a regularized logistic LASSO regression both to the initial dataset and to datasets enriched with additional variables generated by both aforementioned algorithms.

As a measure of the model's improvement in predictive power, the Lift 5% measure is chosen. This measure reflects how well the model fulfils its business requirements in terms of prediction.

The results obtained confirm all three hypotheses formulated above.

The results show that an improvement of up to 25% in the models' predictive power is possible in terms of the Lift 5% criterion. The improvement can be obtained by including variables mentioned in the first two hypotheses in the modeling dataset. Those variables are the result of the proposed algorithms.

In spite of the fact that models enriched with additional predictors are overtrained to some extent, they deliver better predictions on the independent test set used than the standard model does on the training set. Stability analysis proves that overtraining of the models proposed in this thesis can be reduced by limiting the number of predictors included in the models. Such activity contributes to further improvement of the models' predictive power. This confirms the third hypothesis.

Stability analyses conducted in terms of sample selection and modification of the target variable confirm that the solution proposed is stable. The latter analysis endorses the possibility of implementation of both the algorithms presented above in solutions to rare-event problems other than the cases of application fraud defined here.

7. Contribution to knowledge

The contribution of the thesis to knowledge is the introduction of two algorithms that support a solution to the problem of rare-event identification. In the case of the research conducted, the rare event was application fraud. The author succeeded in proving that model enrichment with additional predictors leads to the further improvement of the predictive power of a rare event model and better understanding of the results obtained. The universal construction and assumptions behind both algorithms make it possible to apply them to rare-event problems other than application fraud.

One of the main difficulties the author had to face was limited literature sources in terms of application fraud. Despite the fact that both fraud detection and risk modeling are well-described topics in the literature, modeling of application fraud is commented on quite seldom.

The main motivation for conducting this research was the deficit of modeling methodology that can be exploited in solving the problem of prediction of rare events that are specific to rare, uncommon phenomena. As a result, a new approach was formulated.

In spite of the fact that rare-event modeling is addressed in the literature in many ways, these solutions mostly reduce the bias of estimators due to unbalanced training samples. They do not explain the variability of the phenomena discussed; however, the algorithms introduced here do so.

The method is of general use and can be utilized not only in the case of application fraud modeling but also for other rare events. As the method presented here is not a modeling tool but rather a means of variable selection, it can support research based on a wide range of predictive models – from linear regression to ensemble classifier methods.

The method presented is effective and its performance is stable.

Moreover, model enrichment with additional predictors allows us to better understand the sources of variation within phenomena – in this case, application fraud. Due to the fact that sensitive data was used, this topic is not described in detail here. However, the presented method can be successfully exploited by practitioners of banking credit and applied to solve real decision problems.

References

- Basel Committee on Banking Supervision (2006), International Convergence of Capital Measurement and Capital Standards a Revised Framework, Bank for International Settlements, Basel
- Cruz M. (2002), Modeling, Measuring and Hedging Operational Risk, J. Wiley & Sons, New York
- Dorfleitner G., Jahnes H. (2014), What factors drive personal loan fraud? Evidence from Germany, Review of Managerial Science, 1/8, 89-119
- Farvaresh H., Sepehri M. (2011), A data mining framework for detecting subscription fraud in telecommunication, Engineering Applications of Artificial Intelligence, Volume 24, Issue 1, 182–194
- Hand D.J. (2007), Mining Personal Banking Data to Detect Fraud, Selected Contributions in Data Analysis and Classification, 377–386
- Hartmann-Wendels T., Mählmann T., Versen T. (2009), Determinants of banks' risk exposure to new account fraud – Evidence from Germany, Journal of Banking & Finance, 33:347–357
- Hawkins D. (1980), Identification of Outliers, Chapman and Hall Hawkins, London
- Iwanicz-Drozdowska M. (2012), Zarządzanie ryzykiem bankowym (editor and co-author), Wydawnictwo Poltext, Warszawa
- Kim Y., Sohn S. (2012), Stock fraud detection using peer group analysis, Expert Systems with Applications, 39:8986–8992
- Krivko M. (2010), A hybrid model for plastic card fraud detection systems, Expert Systems with Applications, 37:6070–6076
- Mählmann T. (2010), On the correlation between fraud and default risk, Zeitschrift für Betriebswirtschaft, December, Volume 80, Issue 12, 1325–1352
- Rokach L. (2010), Ensemble-based classifiers, Artificial Intelligence Review, 33:1–39
- Sanusi Z., Rameli M., Isa Y. (2015), Fraud Schemes in the Banking Institutions: Prevention Measures to Avoid Severe Financial Loss, Procedia Economics and Finance, 28:107–113
- Thornton D., Capelleveen G., Poel M., Hillegersberg J., Müller R. (2014), Outlier-based Health Insurance Fraud Detection for U.S. Medicaid Data, 16th International

Conference on Enterprise Information Systems, ICEIS 2014, 27-30 April 2014, Lisbon, Portugal, 684–694

- Tibshirani T. (1996), Regression Shrinkage and Selection via the Lasso, Journal of the Royal Statistical Society. Series B (Methodological), 58:267–288
- Van Vlasselaer V., Eliassi-Rad T., Akoglu L., Snoeck M., Baesens B. (2015), Gotcha! Network-based Fraud Detection for Social Security Fraud, Management Science (submitted)

Damien Paolaggi